Reproducibility Studies and Interlaboratory Concordance for Assays of Serum Hormone Levels: Estrone, Estradiol, Estrone Sulfate, and Progesterone

Mitchell H. Gail,[1] Thomas R. Fears, Robert N. Hoover, Donald W. Chandler, Jennifer L. Donaldson, Marianne B. Hyer, David Pee, Winifred V. Ricker, Pentti K. Siiteri, Frank Z. Stanczyk, Jimmie B. Vaught, and Regina G. Ziegler

Biostatistics Branch [M. H. G., T. R. F., J. L. D.] and Environmental Epidemiology Branch [R. N. H., P. K. S., R. G. Z.], National Cancer Institute, NIH, Bethesda, Maryland 20892; Endocrine Sciences, Calabasas, California 91304 [D. W. C.]; Information Management Services, Rockville, Maryland 20852 [M. B. H., D. P., W. V. R.]; Woman's and Children's Hospital, Los Angeles, California 90033 [F. Z. S.]; and Microbiological Associates, Rockville, Maryland 20850 [J. B. V.]

Abstract

We conducted studies to measure sources of assay variability for estrone, estradiol, estrone sulfate, and progesterone for postmenopausal women ($n = 5$) and for women in the mid-follicular ($n = 5$) and mid-luteal ($n = 5$) phases of the menstrual cycle. A single blood sample from each woman was divided into 2.5-ml aliquots and stored at −70°C, and sets of two aliquots were sent at monthly intervals to each of three laboratories (four for progesterone). Each aliquot was analyzed in duplicate. Thus, within each menstrual category, we were able to estimate the components of variance due to variation among women, variation among aliquots, variation among duplicate measurements, and variation among the 4 analysis days. Using the logarithm of assay measurements, we estimated the percentage of variance attributable to variation among women in each menstrual category, 100 $\hat{\rho}$, where $\hat{\rho}$ is the estimated intraclass correlation. For each assay, 100 $\hat{\rho}$ exceeded 90% for mid-follicular and mid-luteal women. For postmenopausal women, values of 100 $\hat{\rho}$ exceed 84% for estrone in two laboratories. Values of 100 $\hat{\rho}$ were lower for progesterone in postmenopausal women, although a value of 84% was estimated from one laboratory. These studies indicate that estrogen assays over a period of 3 months permit reliable comparisons among women in a given menstrual category. Progesterone measurements are likewise reliable for women in the mid-follicular and mid-luteal phases but somewhat less satisfactory for postmenopausal women. These assessments of variability pertain only to laboratory techniques and do not allow for secular variation in intra-woman hormone levels. Moreover, although these measurements tend to be reliable enough for making comparisons among women, estimates of coefficients of variation for estrogens are about 10% for mid-follicular and mid-luteal phase women and about 11–20% for postmenopausal women. Coefficients of variation for progesterone are about 10% for mid-luteal, 20% for mid-follicular, and 30% for postmenopausal women.

Introduction

NCI[2] has sponsored and is planning several field studies to evaluate associations between serum hormone levels and risks of various cancers, such as breast cancer, endometrial cancer, and prostate cancer. The success of such studies depends on the reproducibility and accuracy of hormone assays as performed by laboratories with the capacity to perform large numbers of tests. Concerns have been raised that the degree of variability in assay results is so great as to degrade the power of studies to detect associations between hormone levels and cancer risks (1, 2). For these reasons, the NCI has conducted a feasibility trial to determine the reproducibility of assay results on the same day and across time in four laboratories with the potential to perform these assays for large-scale epidemiological surveys. Epidemiological field studies may require that samples be analyzed over a period of months or years. For example, samples may be mailed for analysis during the course of a long study, or limitations on laboratory capacity may necessitate carrying out analyses over a period of months. For this reason, we have estimated assay variability over time using monthly measurements over 3 months. The present report deals with the four hormones estrone, estradiol, estrone sulfate, and progesterone.

Because these hormones are influenced by menopausal status and by phase of the menstrual cycle (postmenopausal, mid-follicular phase, and mid-luteal phase), and because some epidemiological studies focus on women in a particular menopausal status or menstrual phase, reproducibility data are presented separately within these three categories.

Materials and Methods

Collection and Distribution of Samples. Plasma for the hormone assays was collected from 15 (1 black and 14 white) volunteers working at the NCI. Mid-follicular phase bloods were collected 6–10 days after start of menses from five women with regular cycles (mean age, 40 years). Mid-luteal

Received 12/12/95; revised 4/2/96; accepted 4/3/96.
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] To whom requests for reprints should be addressed, at Biostatistics Branch, National Cancer Institute, Executive Plaza North, Room 431, MSC 7368, 6130 Executive Boulevard, Bethesda, MD 20892-7368.

[2] The abbreviations used are: NCI, National Cancer Institute; CV, coefficient of variation.

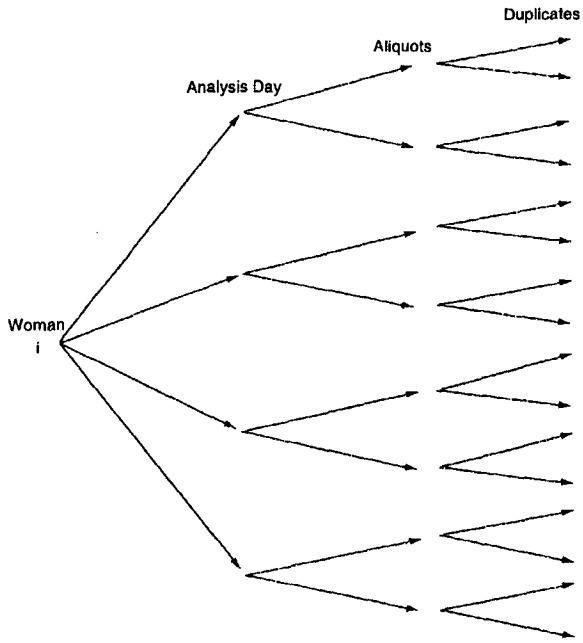


Fig. 1. Measurements on specimens from a single subject.

phase bloods (including one from the black volunteer) were collected 4–6 days prior to the estimated start of next menses from five women with regular cycles (mean age, 39 years), and subsequent follow-up confirmed that menses began 4–6 days after the blood was collected. The timing of the blood draws was confirmed with the date of the subsequent menses. Five women had experienced natural menopause, with at least 3 years since their last cycle (mean age, 56 years). No women were currently taking exogenous estrogens.

Approximately 500 ml of blood was drawn from each woman into a bag containing 750 mg EDTA, equivalent to the EDTA concentration in lavender-top vacutainers. Plasma was separated by centrifugation and stored at 4°C. Within 24 h, the plasma was mixed carefully and aliquotted into 2.5-ml portions, which were stored at −70°C.

Each participating laboratory received four batches of samples, with one batch to be assayed at the beginning of each of 4 consecutive months. Each batch contained two aliquots from each of the 15 subjects. The identifying numbers for the 30 samples within each batch were assigned randomly, separately for each batch. Laboratory personnel were told only whether a sample was from a pre- or postmenopausal woman. Each aliquot was assayed in duplicate. This study design is depicted in Fig. 1 for a single woman and a single laboratory.

Laboratory Methods. Four laboratories, two academic and two commercial, recognized for their skill and experience in measuring endogenous hormones, were invited to participate in this study. Each laboratory was asked to use their standard assay procedures and to perform only those assays with which they had experience. The term "sensitivity" used below refers to the lowest mean value that a laboratory will report for replicate measurements. Sensitivity thus refers to a lower threshold value for reporting.

Laboratory 1. Estradiol and estrone were measured by ethyl acetate:hexane extraction, chromatography on celite, and specific RIA (3–7). Estrone sulfate was measured by ethyl acetate: hexane extraction of unconjugated estrone, overnight hydrolysis of the sulfate conjugate in the aqueous phase, ethyl acetate: hexane extraction of the hydrolyzed compound, and specific RIA for estrone (7–10). Progesterone was measured by ethyl acetate:hexane extraction and specific RIA (11–14). Sensitivity of the assays reported by this laboratory was 2 pg/ml for estradiol, 10 pg/ml for estrone, 50 pg/ml for estrone sulfate, and 5 ng/dl for progesterone. This laboratory reported intra-assay CVs for medium-range quality control pools of 6.2% for estradiol, 8.7% for estrone, 7.5% for estrone sulfate, and 8.8% for progesterone. Interassay CVs were 7.5% for estradiol, 11% for estrone, 9.6% for estrone sulfate, and 10% for progesterone.

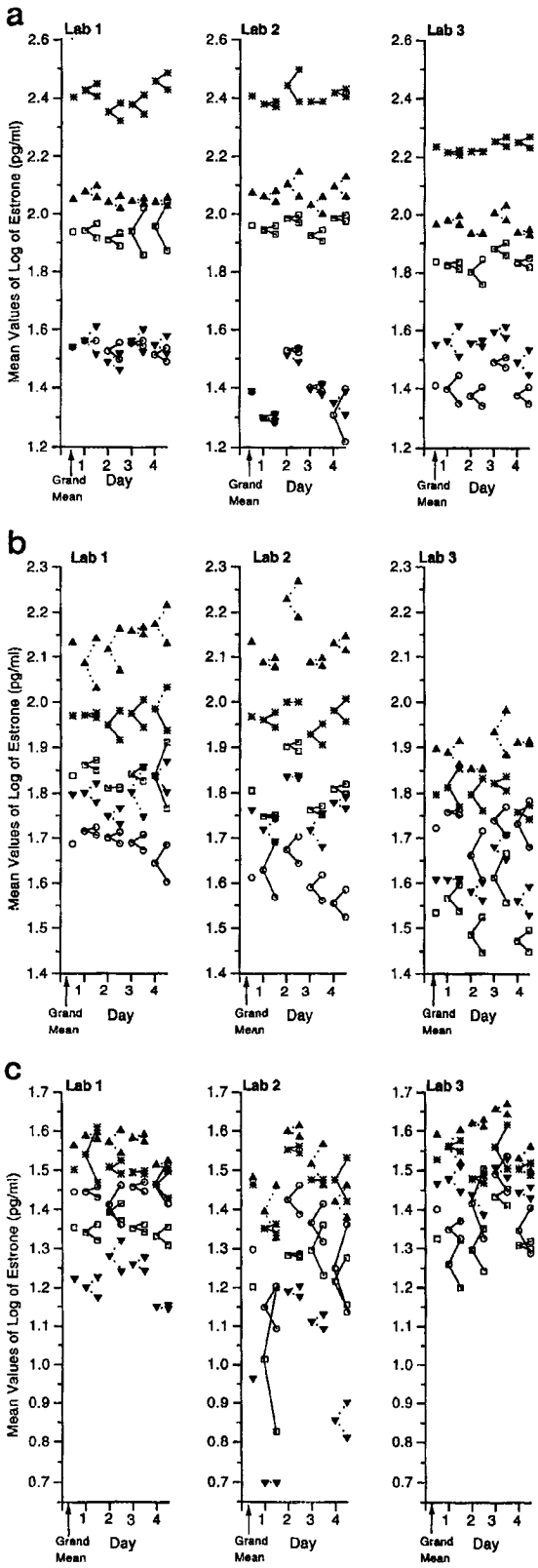
Laboratory 2. Estradiol and estrone were measured by ethyl acetate:hexane extraction, chromatography on Sephadex LH-20, and specific RIA using a modification of the procedure of Wu and Lundy (15). Estrone sulfate was measured by ethyl acetate:hexane extraction of unconjugated estrone, followed by overnight hydrolysis of the sulfate-conjugated compounds in the aqueous phase. The resulting estrone was extracted with ethyl acetate:hexane, followed by chromatography on micro-Sephadex LH-20 and the RIA for estrone. Progesterone was measured by ethyl acetate:hexane extraction and a specific RIA developed in Laboratory 3. Sensitivity of the assays reported by this laboratory was 5 pg/ml for estradiol, 5 pg/ml for estrone, 300 pg/ml for estrone sulfate, and 10 ng/dl for progesterone. This laboratory reported interassay CVs for medium-range quality control pools of 10.8% for estradiol, 9.2% for estrone, 10.5% for estrone sulfate, and 10% for progesterone.

Laboratory 3. Estradiol was measured by ethyl acetate:hexane extraction and a double antibody RIA kit (Pantex, Santa Monica, CA). A modification of the procedure recommended in the package insert was used (16). Estrone and progesterone were measured directly in plasma using RIA kits from Diagnostic Systems Laboratories (Webster, TX) without prior extraction or chromatography. Estrone sulfate was measured by ethyl acetate:hexane extraction of unconjugated estrone, followed by overnight hydrolysis of the sulfate-conjugated compounds in the aqueous phase and the RIA for estrone (8). Sensitivity of the assays reported by this laboratory was 8 pg/ml for estradiol, 15 pg/ml for estrone, 160 pg/ml for estrone sulfate, and 10 ng/dl for progesterone. This laboratory reported intra-assay CVs of 9.2% for estradiol, 7.4% for estrone, 7.5% for estrone sulfate, and 7.5% for progesterone. Interassay CVs were 12.5% for estradiol, 7.7% for estrone, 11% for estrone sulfate, and 10% for progesterone.

Laboratory 4. Progesterone was measured using a Coat-A-Count RIA kit (Diagnostics Production Corporation, Los Angeles, CA), without prior extraction or chromatography. Sensitivity of the assay, as specified in the kit documentation, was 10 ng/dl. This laboratory reported an intra-assay CV for medium-range quality control pools of 8.5% for progesterone. The interassay CV was 8.3%.

Statistical Methods. Data were analyzed on the logarithmic scale (base 10), because this transformation reduced the dependence of the SD of the response on the mean response. Another rationale for this transformation is that studies of cancer associations will typically regress log relative risk on log (hormone) assay levels.

Graphs depict the grand mean overall observations for each study subject, the mean over aliquots and replicates for a subject on each of 4 days, and the mean over replicates for each aliquot (Figs. 2–5). From these figures one can gauge stability of assay results over time, the magnitudes of various sources of



assay variability in relation to the variability among women, and the degree of concordance of results among laboratories analyzing samples drawn from the same women at one point in time. More detail on how to read these graphs is given in "Results."

To estimate components of variance associated with variation among women (σ^2_a), variation among days of analysis (σ^2_b), variation among aliquots on a given day (σ^2_c), and variation among replicate measurements for a given aliquot (σ^2), we performed a nested ANOVA separately for each group of women classified by menstrual phase. Letting y_{ijkl} denote the logarithm (base 10) of the hormone measurement for woman i (i = 1, 2, 3, 4, or 5) at analysis day j(i) (1, 2, 3, or 4), on aliquot k(ij) (1 or 2) and replicate l(ijk) (1 or 2), we define the statistical model

$$y_{ijkl} = \mu + a_i + b_{j(i)} + c_{k(ij)} + \epsilon_{l(ijk)}. \quad (A)$$

In model A, a_i, $b_{j(i)}$, $c_{k(ij)}$ and $\epsilon_{l(ijk)}$ are independent normal variates with mean 0 and respective variances σ^2_a, σ^2_b, σ^2_c and σ^2, and μ is an overall mean. Restricted maximum likelihood estimates of the variance components were obtained using the SAS procedure for estimating variance components in a general linear model, PROC VARCOMP (17). This procedure also yields an estimate of the variances and covariances of the estimated variance components. The four required ordered SAS statements are: PROC VARCOMP METHOD = REML; CLASSES WOMAN DAY ALIQUOT; MODEL LOG_E1 = WOMAN DAY(WOMAN) ALIQUOT(WOMAN DAY);.

Under model A, the intraclass correlation between two measurements on different days from a given individual is $\rho = \sigma^2_a/(\sigma^2_a + \sigma^2_b + \sigma^2_c + \sigma^2)$. The intraclass correlation is high when the variance component associated with women (σ^2_a) greatly exceeds the sum of the variance components associated with the assay. Using the "δ method" (18) and estimates of the variances and covariances of the estimated variance components, we obtained an estimate of the SE of $\hat{\rho}$.

Spearman rank correlations were used to estimate concordance of grand mean results among laboratories.

Results

Estrone. Fig. 2*a* depicts the results for log(estrone) among mid-follicular phase women. The leftmost symbols for Laboratory 1 (Fig. 2*a, Lab 1*) are grand means of the 4 × 2 × 2 = 16 measurements of log(estrone) for each of the five women. Note the large differences among these women. The next column of symbols corresponds to the means of the 2 × 2 = 4 measurements for each woman on analysis day 1, and attached to these symbols are means of replicates for the two separate aliquots on that day. Fig. 2*a* thus allows one to assess not only variation among individuals but variation among analysis days for a given individual and among aliquots on a given day. Variability among replicates is not depicted. Fig. 2*a* also allows one to compare Laboratories 1, 2, and 3 not only as to mean levels of response, but also as to variability. The same symbols are used to identify an individual across laboratories.

Fig. 2. Estrone measurements in mid-follicular phase (*a*), mid-luteal phase (*b*), and postmenopausal (*c*) women. The leftmost symbols for Laboratory 1 (*Lab 1*) are grand means of the 4 × 2 × 2 = 16 measurements of log(estrone) for each of the five women. The next column of symbols corresponds to the means of the 2 × 2 = 4 measurements for each woman on analysis day 1, and attached to these symbols are means of replicates for the two separate aliquots on that day. The same symbols (*, ▲, ▼, □, ○) are used to identify an individual across laboratories.

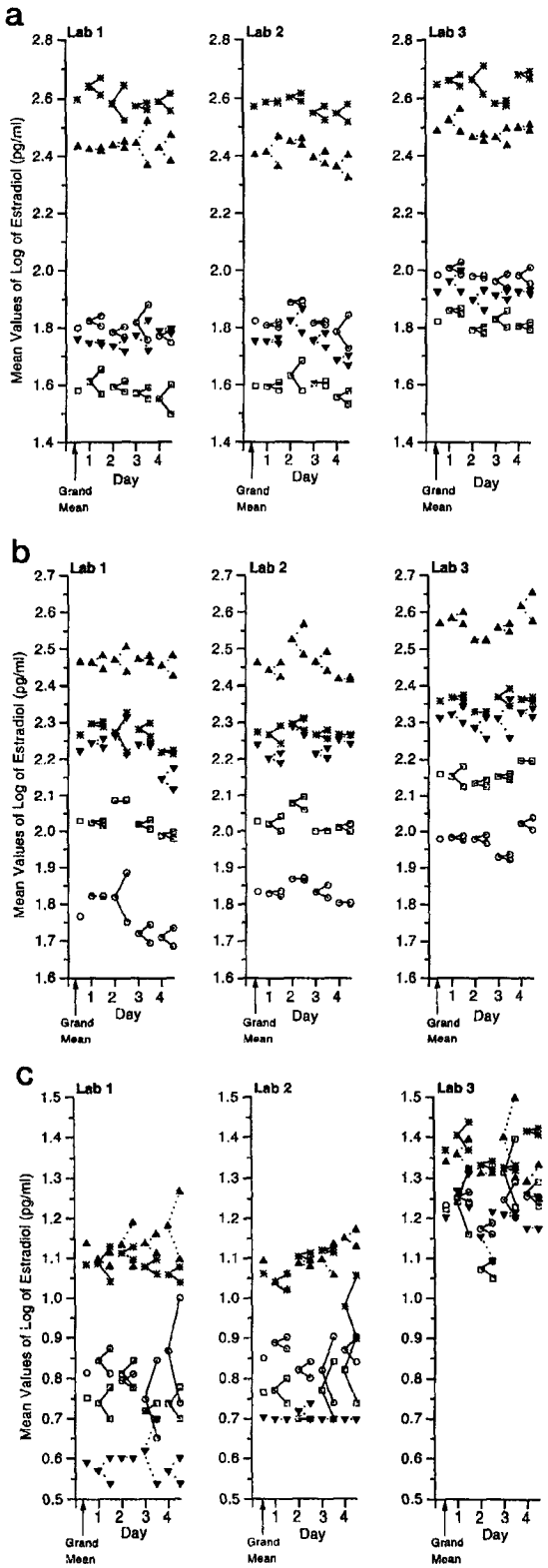


Fig. 3. Estradiol measurements. See Fig. 2 for details.

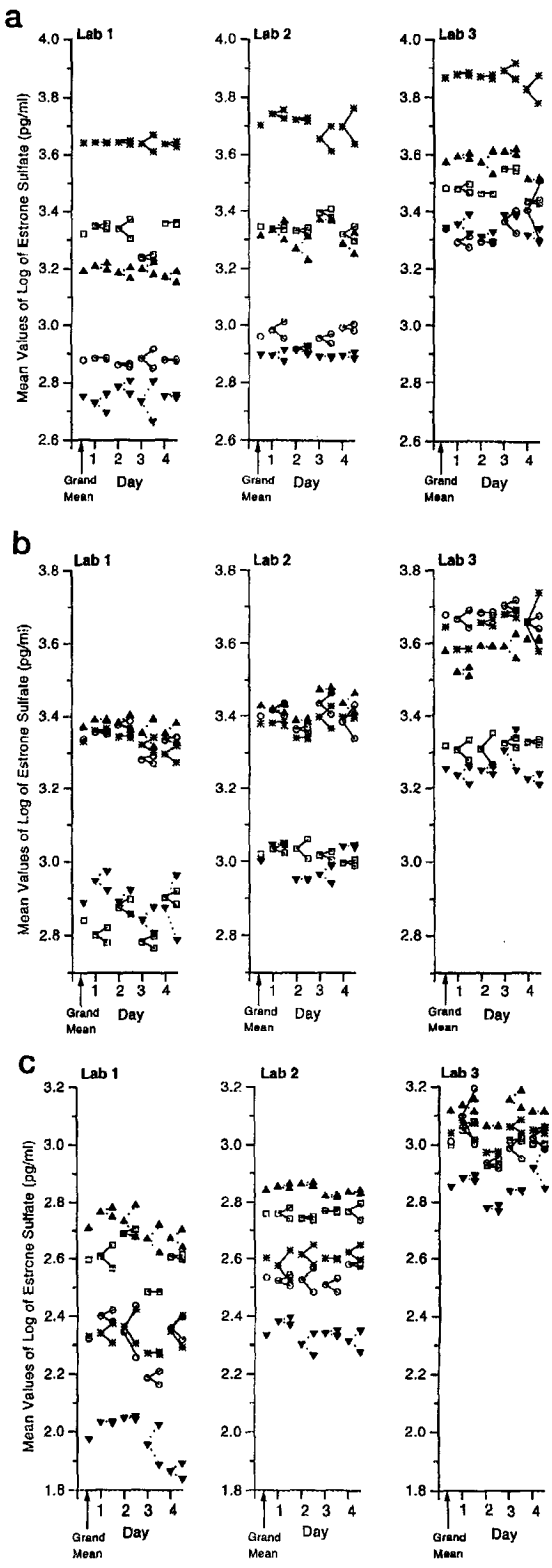


Fig. 4. Estrone sulfate measurements. See Fig. 2 for details.

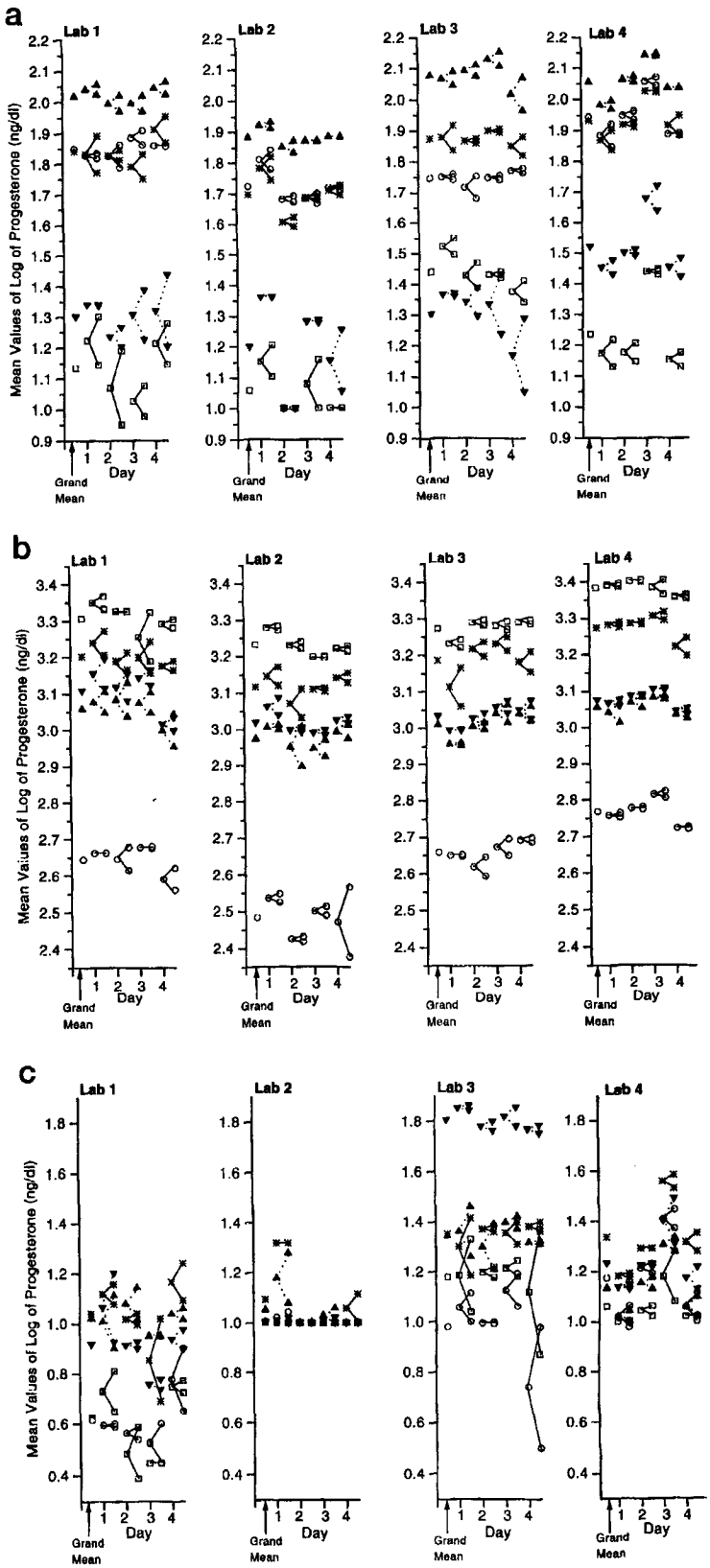


Fig. 5. Progesterone measurements. See Fig. 2 for details.

There is little evidence for time trends in any menopausal phase (Fig. 2 *a–c*). Results are separated and rather consistent across laboratories for mid-follicular women (Fig. 2*a*), but the separation among women is less distinct for mid-luteal women (Fig. 2*b*) and even less well defined for postmenopausal women, whose absolute values are also much lower (Fig. 2*c*).

The percentages of variation attributable to variation among mid-follicular women (100 $\hat{\rho}$) are 97.9, 96.5, and 97.0%, respectively, for the three laboratories (Table 1*a*). These percentages fall to 91.2, 90.3, and 81.9%, respectively, for mid-luteal women and to 83.6, 61.1, and 62.3% for postmenopausal women. The SEs indicate that the estimates of percentage of variation (100 $\hat{\rho}$) are known with good precision except for Laboratories 2 and 3 in postmenopausal women, although some estimates of variability among women, $\hat{\sigma}_a$, are not very precise.

Correlations of the ranks of the grand means are perfect between Laboratories 1 and 2 within menstrual phase (Table 2) and high (0.99) overall. The ranks from Laboratory 3 do not agree as well with Laboratories 1 and 2, especially among mid-luteal and postmenopausal women.

The mean levels of $\log_{10}$ (estrone) are higher in Laboratory 1 than in the other two Laboratories (Table 3).

Estradiol. There are no consistent time trends in these measurements (Fig. 3, *a–c*). Mid-follicular women are fairly well separated (Fig. 3*a*), as are mid-luteal women (Fig. 3*b*). However, there are many overlapping measurements among postmenopausal women (Fig. 3*c*), who have much lower estradiol measurements than premenopausal women. A lower threshold of sensitivity at 0.7 (assay level $10^{0.7} = 5.0$ pg/ml) is seen for Laboratory 2 among postmenopausal women (Fig. 3*c*).

The percentages of total variation attributable to variation among women (100 $\hat{\rho}$) were 98.7, 98.4, and 98.6%, respectively, for Laboratories 1, 2, and 3 for mid-follicular women (Table 1*b*). For mid-luteal women, these estimates were 96.0, 97.3, and 96.2%, respectively, and for postmenopausal women, the estimates fell to 91.3, 84.9, and 41.5%. Except for postmenopausal women in Laboratory 3, the estimates of percentage of variation (100 $\hat{\rho}$) have good precision, as indicated by relatively small SEs.

Concordance of ranks of women in a given menopausal or menstrual phase was high among all laboratories (Table 2), and perfect between Laboratories 1 and 2. Laboratory 3 had higher average values (Table 3).

Estrone Sulfate. No definite time trends are evident (Fig. 4, *a–c*), although there is a suggestion of a decreasing trend in Laboratory 1 for some mid-luteal (Fig. 4*b*) and postmenopausal (Fig. 4*c*) subjects. There is some overlap of measurements among subjects in all three menstrual phases.

The percentages of total variation attributable to variation among mid-follicular women (100$\hat{\rho}$) were 98.5, 98.1, and 93.6% respectively, for Laboratories 1, 2, and 3 (Table 1*c*). The corresponding percentages for mid-luteal women were 96.1, 96.1, and 94.6%, and for postmenopausal women the percentages were 90.0, 96.5, and 65.7%. Except for postmenopausal women in Laboratory 3, these estimates of 100 $\hat{\rho}$ have good precision.

The ranks of the subjects' mean responses were highly correlated (0.90–1.00) between Laboratories 1 and 2 (Table 2). Correlations with Laboratory 3 were lower and ranged from 0.60 to 0.80 within menstrual phases. Mean levels were significantly different among the three laboratories (Table 3).

Progesterone. There are no obvious time trends (Fig. 5, *a–c*). Laboratory 2 has a lower threshold of sensitivity of $10^{1.0} = 10$ ng/dl, which affects results for postmenopausal women (Fig. 5*c*).

The percentages of total variation attributable to variation among mid-follicular women (100$\hat{\rho}$) were 95.7, 92.7, 92.1, and 92.4%, respectively, for Laboratories 1, 2, 3, and 4 (Table 1*d*). Corresponding percentages were 95.0, 97.4, 96.3, and 98.0% for mid-luteal women and 62.2, 1.8, 80.8, and 23.5% for postmenopausal women. Except for postmenopausal women, these estimates of 100 $\hat{\rho}$ have good precision. The extremely low value of 1.8% in Laboratory 2 among postmenopausal women reflects the impact of setting all values at or below 10 ng/dl to that threshold level.

Correlations among ranks of subjects' means are high among all four laboratories for mid-follicular and mid-luteal women, but interlaboratory correlations are poor among postmenopausal women (Table 2). Laboratories 1 and 2 yielded mean values lower than Laboratories 3 and 4 (Table 3).

Approximate CVs. CVs are usually estimated by repeatedly assaying samples from a single large pool of analyte and dividing the SD of the measurements by the mean value. The CVs typically depend on the mean concentration in the pool, with larger CVs often associated with smaller mean values. The variance of the natural logarithm of a hormone level is, by the δ method (18), roughly the square of the CV. Because the SD of the natural logarithm of an assay value is $\log_e(10) = 2.303$ times the SD of the logarithm to the base 10 of the assay value, we can approximate the CV (in percentage) from the data in Table 1, *a–d*, by

$$100 \times 2.303 \times (\hat{\sigma}^2_b + \hat{\sigma}^2_c + \hat{\sigma}^2/2)^{1/2}.$$

For example, from Table 1a, the CV for estrone in mid-follicular women is approximately $100 \times 2.303 \times (0 + 0.042^2 + (0.034)^2/2)^{1/2} = 11.1\%$, as in Table 4. The divisor 2 in $\hat{\sigma}^2/2$ arises because we are estimating the assay variability for a randomly selected day and aliquot based on the mean of duplicate assay measurements.

Based on these methods, we estimate that the CVs for estrogens are near 10% for mid-follicular and mid-luteal phase women, except that Laboratory 2 has somewhat higher CVs for estrone. CVs tend to be between 11 and 20% for postmenopausal women.

For progesterone, CVs are near 10% for mid-luteal phase women, who have the highest progesterone levels, near 20% for mid-follicular phase women, who have the next highest progesterone levels, and near 30% for postmenopausal women.

Discussion

This study provides data on components of variability in estrogen and progesterone assay results and allows a comparison between biological variability among women in a given menstrual phase and other sources of variability including monthly variation in assay procedures, aliquot variation, and replication error on a given aliquot and day.

The estrogen data for mid-follicular and mid-luteal women show that most of the variability (more than 90%) is due to variability among women. Even for postmenopausal women, among whom estrogen levels are much lower, the proportion of variability attributable to variation among women exceeds 84% for estradiol and estrone sulfate (except for Laboratory 3). These data indicate that a single estrone, estradiol, or estrone sulfate measurement can discriminate among premenopausal women in the same menstrual phase and that a single estradiol or estrone sulfate measurement can discriminate among post-

Table 1 Estimated square roots of variance components and percentage of variation for logarithms (base 10) of estrone (*a*), estradiol (*b*), estrone sulfate (*c*), and progesterone (*d*)

Components	Laboratory 1		Laboratory 2		Laboratory 3		Laboratory 4	
	Square root of components (SE)	% variation[a] (SE)	Square root of components (SE)	% variation (SE)	Square root of components (SE)	% variation (SE)	Square root of components (SE)	% variation (SE)
				a. log(estrone)[b]				
Mid-follicular phase women								
Subject (σ^2_a)	0.366 (0.130)	97.9 (1.5)	0.446 (0.159)	96.5 (2.5)	0.327 (0.116)	97.0 (2.2)		
Analysis day (σ^2_b)	0	0.0	0.060 (0.014)	1.8	0.029 (0.010)	0.8		
Aliquot (σ^2_c)	0.042 (0.007)	1.3	0.017 (0.021)	0.1	0.016 (0.016)	0.2		
Replication (σ^2)	0.034 (0.004)	0.9	0.057 (0.006)	1.6	0.047 (0.005)	2.0		
Mid-luteal phase women								
Subject (σ^2_a)	0.171 (0.061)	91.2 (6.0)	0.198 (0.071)	90.3 (6.8)	0.143 (0.052)	81.9 (11.2)		
Analysis day (σ^2_b)	0	0.0	0.051 (0.012)	5.9	0.033 (0.013)	4.4		
Aliquot (σ^2_c)	0.038 (0.007)	4.5	0.028 (0.007)	1.8	0.029 (0.013)	3.4		
Replication (σ^2)	0.037 (0.004)	4.4	0.029 (0.003)	2.0	0.051 (0.006)	10.3		
Postmenopausal women								
Subject (σ^2_a)	0.132 (0.048)	83.6 (10.2)	0.200 (0.080)	61.1 (20.5)	0.100 (0.038)	62.3 (18.8)		
Analysis day (σ^2_b)	0.023 (0.012)	2.6	0.127 (0.029)	24.3	0.041 (0.016)	10.3		
Aliquot (σ^2_c)	0.024 (0.013)	2.7	0.077 (0.016)	8.9	0.045 (0.012)	12.3		
Replication (σ^2)	0.048 (0.005)	11.1	0.061 (0.007)	5.7	0.049 (0.006)	15.1		
				b. log(estradiol)[c]				
Mid-follicular phase women								
Subject (σ^2_a)	0.450 (0.160)	98.7 (0.9)	0.430 (0.152)	98.4 (1.2)	0.369 (0.131)	98.6 (1.0)		
Analysis day (σ^2_b)	0	0.0	0.029 (0.011)	0.5	0.020 (0.010)	0.3		
Aliquot (σ^2_c)	0.039 (0.006)	0.7	0.034 (0.008)	0.6	0.026 (0.007)	0.5		
Replication (σ^2)	0.033 (0.004)	0.5	0.031 (0.003)	0.5	0.029 (0.003)	0.6		
Mid-luteal phase women								
Subject (σ^2_a)	0.264 (0.094)	96.0 (2.9)	0.242 (0.086)	97.3 (2.0)	0.222 (0.079)	96.2 (2.7)		
Analysis day (σ^2_b)	0.034 (0.011)	1.6	0.029 (0.007)	1.4	0.021 (0.008)	0.9		
Aliquot (σ^2_c)	0.035 (0.007)	1.7	0.021 (0.005)	0.7	0.013 (0.012)	0.3		
Replication (σ^2)	0.024 (0.003)	0.8	0.019 (0.002)	0.6	0.036 (0.004)	2.6		
Postmenopausal women								
Subject (σ^2_a)	0.229 (0.082)	91.3 (5.9)	0.174 (0.062)	84.9 (9.5)	0.069 (0.029)	41.5 (22.0)		
Analysis day (σ^2_b)	0	0.0	0.022 (0.020)	1.4	0.045 (0.017)	18.1		
Aliquot (σ^2_c)	0.060 (0.009)	6.3	0.036 (0.015)	3.6	0.052 (0.011)	23.8		
Replication (σ^2)	0.036 (0.004)	2.3	0.060 (0.007)	10.1	0.043 (0.005)	16.5		
				c. log(estrone sulfate)[d]				
Mid-follicular phase women								
Subject (σ^2_a)	0.356 (0.126)	98.5 (1.1)	0.326 (0.116)	98.1 (1.4)	0.216 (0.077)	93.6 (4.5)		
Analysis day (σ^2_b)	0.016 (0.011)	0.2	0.022 (0.010)	0.5	0.028 (0.013)	1.6		
Aliquot (σ^2_c)	0.023 (0.008)	0.4	0.033 (0.006)	1.0	0.040 (0.008)	3.3		
Replication (σ^2)	0.034 (0.004)	0.9	0.022 (0.003)	0.4	0.027 (0.003)	1.5		
Mid-luteal phase women								
Subject (σ^2_a)	0.264 (0.094)	96.1 (2.8)	0.215 (0.076)	96.1 (2.8)	0.194 (0.069)	94.6 (3.8)		
Analysis day (σ^2_b)	0.029 (0.011)	1.2	0.029 (0.008)	1.7	0.015 (0.015)	0.6		
Aliquot (σ^2_c)	0.034 (0.007)	1.6	0.018 (0.006)	0.7	0.035 (0.007)	3.1		
Replication (σ^2)	0.028 (0.003)	1.1	0.026 (0.003)	1.5	0.027 (0.003)	1.8		
Postmenopausal women								
Subject (σ^2_a)	0.282 (0.101)	90.0 (6.8)	0.198 (0.070)	96.5 (2.5)	0.091 (0.035)	65.7 (18.6)		
Analysis day (σ^2_b)	0.061 (0.017)	4.3	0.006 (0.024)	0.1	0.044 (0.013)	15.4		
Aliquot (σ^2_c)	0.037 (0.015)	1.5	0.031 (0.006)	2.3	0.041 (0.008)	13.4		
Replication (σ^2)	0.062 (0.007)	4.3	0.021 (0.002)	1.1	0.026 (0.003)	5.5		
				d. log(progesterone)[e,f]				
Mid-follicular phase women								
Subject (σ^2_a)	0.386 (0.137)	95.7 (3.1)	0.357 (0.128)	92.7 (5.3)	0.315 (0.112)	92.1 (5.4)	0.344 (0.124)	92.4 (5.6)
Analysis day (σ^2_b)	0.015 (0.050)	0.1	0.083 (0.018)	5.0	0.034 (0.019)	1.0	0.092 (0.018)	6.7
Aliquot (σ^2_c)	0.073 (0.013)	3.5	0.038 (0.010)	1.0	0.012 (0.061)	0.1	0.031 (0.005)	0.8
Replication (σ^2)	0.034 (0.004)	0.7	0.041 (0.005)	1.2	0.085 (0.010)	6.7	0.013 (0.001)	0.1
Mid-luteal phase women								
Subject (σ^2_a)	0.252 (0.090)	95.0 (3.6)	0.286 (0.102)	97.4 (1.9)	0.235 (0.084)	96.3 (2.7)	0.236 (0.084)	98.0 (1.5)
Analysis day (σ^2_b)	0.030 (0.013)	1.4	0.022 (0.012)	0.6	0.031 (0.008)	1.7	0.027 (0.006)	1.3
Aliquot (σ^2_c)	0.039 (0.008)	2.3	0.039 (0.007)	1.8	0.021 (0.006)	0.7	0.011 (0.004)	0.2
Replication (σ^2)	0.030 (0.003)	1.4	0.014 (0.002)	0.2	0.027 (0.003)	1.3	0.017 (0.002)	0.5
Postmenopausal women								
Subject (σ^2_a)	0.203 (0.078)	62.2 (19.0)	0.012 (0.055)	1.8 (17.2)	0.302 (0.109)	80.8 (11.6)	0.079 (0.050)	23.5 (24.6)
Analysis day (σ^2_b)	0.091 (0.031)	12.5	0.073 (0.015)	71.1	0.000	0.0	0.130 (0.026)	63.4
Aliquot (σ^2_c)	0.083 (0.024)	10.5	0.030 (0.008)	12.3	0.108 (0.019)	10.3	0.055 (0.009)	11.2
Replication (σ^2)	0.099 (0.011)	14.8	0.033 (0.004)	14.8	0.100 (0.011)	8.8	0.022 (0.002)	1.9

[a] Percentage of variation is 100 times the ratio of a given variance component to the sum of the variance components.
[b] Units are $\log_{10}$(estrone), where estrone is in pg/ml.
[c] Units are $\log_{10}$(estradiol), where estradiol is in pg/ml.
[d] Units are $\log_{10}$(estrone sulfate), where estrone sulfate is in pg/ml.
[e] Units are $\log_{10}$(progesterone), where progesterone is in ng/dl.
[f] An entire aliquot was missing for one individual. Missing values were estimated by least squares, and degrees of freedom were adjusted.

Table 2 Spearman correlations among the ranks of the grand mean assay values for women in various menstrual phases[a]

	Laboratory 2				Laboratory 3				Laboratory 4
	Estrone	Estradiol	Estrone sulfate	Progesterone	Estrone	Estradiol	Estrone sulfate	Progesterone	Progesterone
Mid-follicular									
Laboratory 1	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.80	1.00
Laboratory 2					1.00	1.00	0.80	0.80	1.00
Laboratory 3									0.80
Mid-luteal									
Laboratory 1	1.00	1.00	0.90	1.00	0.60	1.00	0.60	1.00	1.00
Laboratory 2					0.60	1.00	0.70	1.00	1.00
Laboratory									1.00
Postmenopausal									
Laboratory 1	1.00	1.00	1.00	0.82	0.70	0.90	0.70	0.70	0.50
Laboratory 2					0.70	0.90	0.70	0.41	0.67
Laboratory 3									0.60
All Women									
Laboratory 1	0.989	0.996	0.989	0.990	0.925	0.996	0.943	0.918	0.979
Laboratory 2					0.925	0.996	0.943	0.899	0.988
Laboratory 3									0.914

[a] All correlations above 0.90 are significant at the $P < 0.05$ level.

Table 3 Comparisons among laboratories[a]

	Laboratory 1	Laboratory 2	Laboratory 3	Laboratory 4	Two-sided signed rank *P*
log(estrone)					(laboratory *vs.* laboratory)
Mean	1.732	1.660	1.658		1 *vs.* 2, 0.004
					1 *vs.* 3, 0.048
					2 *vs.* 3, 0.600
Correlations	1.0	0.994[b]	0.921[b]		
		1.0	0.906[b]		
			1.0		
log(estradiol)					
Mean	1.687	1.698	1.907		1 *vs.* 2, 0.5250
					1 *vs.* 3, 0.0001
					2 *vs.* 3, 0.0001
Correlations	1.0	0.998[b]	0.993[b]		
		1.0	0.996[b]		
			1.0		
log(estrone sulfate)					
Mean	2.899	3.036	3.340		1 *vs.* 2, 0.0001
					1 *vs.* 3, 0.0001
					2 *vs.* 3, 0.0001
Correlations	1.0	0.995[b]	0.962[b]		
		1.0	0.973[b]		
			1.0		
log(progesterone)					
Mean	1.846	1.836	2.017	2.013	1 *vs.* 2, 0.229
					1 *vs.* 3, 0.107
					1 *vs.* 4, 0.000
					2 *vs.* 3, 0.000
					2 *vs.* 4, 0.000
					3 *vs.* 4, 0.679
Correlations	1.0	0.989[b]	0.972[b]	0.992[b]	
		1.0	0.975[b]	0.996[b]	
			1.0	0.973[b]	
				1.0	

[a] These comparisons are based on the grand means of all values of the log (assay) for each of 15 women. The grand means and correlations of these grand means are shown.
[b] Indicates $P < 0.001$.

menopausal women. The estrone data from Laboratories 2 and 3 are not as promising for postmenopausal women, although, for Laboratory 1, 84% of the variability was attributable to variation among women. Thus, the estrogen assay performance is good enough to add useful epidemiological information above that provided by menopausal status and menstrual phase.

Progesterone measurements are also quite reliable for discriminating among women in the mid-follicular and mid-luteal phases of the menstrual cycle, with more than 90% of the variability attributable to variation among women. For postmenopausal women, whose progesterone values are lower, the percentage of variation attributable to variation among women

Table 4 Approximate CVs (in percentages)

	Laboratory			
	1	2	3	4
Estrone				
Mid-follicular	11.1	17.1	10.8	
Mid-luteal	10.6	14.2	13.1	
Postmenopausal	10.9	35.6	16.1	
Estradiol				
Mid-follicular	10.5	11.5	8.9	
Mid-luteal	11.9	8.8	8.2	
Postmenopausal	15.0	13.8	17.3	
Estrone sulfate				
Mid-follicular	8.5	9.8	12.1	
Mid-luteal	11.3	8.9	9.8	
Postmenopausal	19.3	8.0	14.5	
Progesterone				
Mid-follicular	18.0	22.1	16.1	22.5
Mid-luteal	12.3	10.6	9.7	7.3
Post-menopausal	32.6	19.0	29.7	32.7

Values shown are $100 \times \log_e(10)\,(\hat{\sigma}^2_b + \hat{\sigma}^2_c + \hat{\sigma}^2/2)^{1/2}$. See text for explanation of relation to the CV.

was 62% for Laboratory 1 and 81% for Laboratory 3, suggesting that more than one measurement might be useful to discriminate among postmenopausal women. The sensitivity threshold of 10 ng/dl used by Laboratories 2 and 4 affected many measurements and severely limited the ability to discriminate among postmenopausal women.

We have emphasized that because hormone levels vary widely among women, compared to laboratory error, it is possible to rank women reliably, in most cases, even within the same menstrual phase. Estimates of CVs indicate, however, that the laboratory component of variation can be large, compared to mean assay levels, especially among postmenopausal women with relatively low hormone levels (Table 4).

The percentage of total variability attributable to variation among women is 100 times the estimated intraclass correlation $\rho = \sigma^2_a/(\sigma^2_a + \sigma^2_b + \sigma^2_c + \sigma^2)$. The intraclass correlation is an important indication of the effect of assay measurement error on study results. One can compare a study with a single measurement on each subject to a study with a large number of measurements on each subject (many days, aliquots, and replicates) in terms of ρ. Regression analyses relating the log relative risk of disease to the log hormone assay level will tend to be attenuated by the factor ρ in the former study, compared to the latter study (19). The number of subjects in the former study needed to have an equivalent power to detect an association as in the latter study is $1/\rho$ times as great as the number of subjects in the latter study. With ρ greater than 0.90, there is little attenuation, and a single measurement per woman provides nearly the same information as many measurements per woman.

There were some variations in the mean assay levels among these laboratories (Table 3), but the correlations of rankings of mean subjects' results among laboratories were good, especially between Laboratories 1 and 2.

Laboratory 1 exhibited relatively high intraclass correlations for all assays and menstrual phases, and Laboratory 2 also yielded high intraclass correlations except for progesterone assays in postmenopausal women, whose values often fell below a sensitivity threshold.

The estimated values of ρ are subject to systematic and random uncertainty, the latter reflected in the relatively large SEs in Table 1, *a–d*, for postmenopausal women. To estimate ρ more accurately and precisely, larger numbers of randomly selected women in each menstrual class would need to be studied, or external information on assay variation among women in each menstrual phase could be used. To estimate ρ precisely, one needs a more precise estimate of the variability among women, σ_a^2, than can be obtained from the study of only a few women. For postmenopausal women, we compared results in Table 1, *a* and *b*, with published data for estrone and estradiol, with care taken to translate results in the literature to estimates of σ_a^2 appropriate to logarithmic measurements (base 10). For estrone, the value $\sigma_a^2 = 0.171^2 = 0.029$ from Laboratory 1 (Table 1*a*) is very near the average laboratory estimate, 0.027, derived from Table 1 in Hankinson *et al.* (2). We calculated the corresponding estimate from the data in Table 2 of Cauley *et al.* (20) as $\{11.9^2/[29.9^2 \times \ln^2(10)]\} \times 0.836 = 0.025$. This formula is based on the δ method approximation (18) to the variance of $\log_{10}$ (estrone), and the factor 0.836 (from our Table 1*a*) was used to convert the total variance of an observation to σ_a^2. The 176 postmenopausal women in the study of Cauley *et al.* (20) were white inhabitants of the metropolitan Pittsburgh area who were participating in a clinical trial to evaluate the effect of walking on postmenopausal bone loss. These results suggest, if anything, that estimates of σ_a^2 for Laboratory 1 may be slightly too small, and that estimates of ρ for Laboratory 1 should be higher for postmenopausal women. For $\log_{10}$ (estradiol), the result $\hat{\sigma}_a^2 = 0.053$ for Laboratory 1 might be compared to derived estimates 0.071, 0.036, and 0.063, respectively, from Refs. 2, 21, and 20. The latter estimate is probably too small, because 52% of the observations were set to a lower threshold of sensitivity, 2.5 pg/ml, when a value fell below threshold. The data from Ref. 21 may have been truncated at the sensitivity threshold of 10 pg/ml, accounting for the smaller estimate of σ_a^2 found. Thus, again, we take these data as indicating that estimates of ρ for Laboratory 1 in Table 1*b* are, if anything, somewhat too small. Additional studies of this type in randomly selected women would be useful to obtain more accurate and precise estimates of ρ.

In principle, one can learn from analyses of components of variance (Table 1, *a–d*) how to efficiently allocate effort to increase the reproducibility of study results. For example, if there were much more variation among aliquots than among replicates or days, one would increase the number of aliquots used per subject. However, given the limited precision of our results and the relatively small variability in assay performance, compared to interindividual variation, we do not plan to pursue this idea.

The estimated components of variance in Table 1 can also be used to plan case-control studies. One can determine the sample sizes needed to reliably detect a given difference, δ, in hormone levels between cases and controls, or, alternatively, to calculate the minimum difference that is reliably detectable with a fixed number of cases and controls. For a two-sided $\alpha = 0.05$ level test, the minimum difference detectable with power 0.9 is the solution to

$$\delta^2 = (\sigma_A^2 + \sigma_B^2 + \sigma_C^2 + \sigma^2/2) \times (1/n_1 + 1/n_2) \times (1.96 + 1.282)^2 \quad \text{(B)}$$

where n_1 is the number of cases, n_2 is the number of controls, $Z_\alpha = 1.96$ is the 97.5th percentile of the standard normal distribution, and $Z_\beta = 1.282$ is the 90th percentile. The quantity $\sigma^2/2$ is used for an experiment in which assay measurements on the same aliquot will be performed in duplicate. For example,

with $n_1 = 218$ cases and $n_2 = 2n_1 = 436$ controls, and using the estimated components of variance for log(estrone) in Table 1*a* for Laboratory 1, we calculated minimum detectable differences 0.099, 0.048, and 0.038, respectively, for mid-follicular, mid-luteal, and postmenopausal women. These differences correspond to percentage increases among cases compared to controls of 26, 12, and 9.1%, respectively. For example $(10^{0.099} - 1) \times 100 = 26\%$. These minimum detectable differences or percentage increases in estrone are determined principally by our estimates of variation among women, $\hat{\sigma}^2_a$, which usually exceed other sources of variation. Because $\hat{\sigma}^2_a$ is imprecisely known, the estimates of minimal detectable differences are also uncertain. It is worth noting that smaller differences are detectable among postmenopausal women than among women in the mid-follicular or mid-luteal menstrual phases, despite the fact that the intraclass correlations are higher in the latter groups. This is because the biological component of variability, represented by $\hat{\sigma}^2_a$, is smallest in postmenopausal women.

We have mentioned that the sensitivity thresholds of 10 ng/dl for progesterone used by Laboratories 2 and 4 limited their ability to discriminate among postmenopausal women. These effects of censoring at a threshold are apparent in Fig. 5*c*. Such censoring also affects the statistical analyses we used and usually leads to smaller estimates of intraclass correlation and CVs than would have been observed in the absence of a threshold. The figures suggest, however, that these distortions would only be appreciable for postmenopausal women for progesterone (Laboratories 2 and 4) and estradiol (Laboratory 2).

One important component of variability was not estimated in this study, namely, the biological variability for a given woman from day to day. Instead, our study used aliquots from a given woman all prepared from blood drawn at one time, so that our estimates σ_b^2 in Table 1 reflect only month-to-month variation in laboratory performance. If one desires to estimate the long-term average hormone level for a woman, and if biological variability from day to day is appreciable, one may need to take several measurements on the same woman over time. For example, Cauley *et al.* (22) estimated another "intraclass correlation" coefficient, namely, the intraclass correlation of measurements on the same woman over time. This quantity will be reduced either when there is substantial within-woman biological variability over time or substantial assay variability over time. For postmenopausal women, they found intraclass correlations over 2 years of 0.56, which was comparable to the intraclass correlation within women of blood pressure measurements. Such data suggest that if the aim of a study is to relate cancer risk to long-term average hormone levels, more than one observation per woman may be more cost efficient than studying larger numbers of women.

The present feasibility study was designed to answer a different question for case-control studies in which a single sample was obtained for each woman, namely, are the laboratory procedures precise enough to permit reliable rankings and comparisons among women in the same menstrual phase. With some qualifications, especially for postmenopausal women, our data give cause to believe such studies are feasible with assays performed over several months, even based on a single aliquot per subject.

References

1. Potischman, N., Falk, R. T., Laiming, V. A., Siiteri, P. K., and Hoover, R. N. Reproducibility of laboratory assays for steroid hormones and sex hormone-binding globulin. Cancer Res., *54:* 5363–5367, 1994.

2. Hankinson, S. E., Manson, J. E., London, S. J., Willett, W. C., and Speizer, F. E. Laboratory reproducibility of endogenous hormone levels in postmenopausal women. Cancer Epidemiol., Biomarkers & Prev., *3:* 51–56, 1994.

3. England, B. G., Niswender, G. D., and Midgley, A. R. Radioimmunoassay of estradiol-17 β without chromatography. J. Clin. Endocrinol. & Metab., *38:* 42–50, 1974.

4. Abraham, G. E., Odell, W. D., Swerdloff, R. S., and Hopper, K. Simultaneous radioimmunoassay of plasma FSH, LH, progesterone, 17-hydroxyprogesterone and estradiol 17-β during the menstrual cycle. J. Clin. Endocrinol. & Metab., *34:* 312–318, 1972.

5. Judd, H. L., Lucas, W. E., and Yen, S. S. Serum 17β-estradiol and estrone levels in postmenopausal women with and without endometrial cancer. J. Clin. Endocrinol. & Metab., *43:* 272–278, 1976.

6. Schneider, G., Kirschner, M. A., Berkowitz, R., and Ertel. N. H. Increased estrogen production in obese men. J. Clin. Endocrinol. & Metab., *48:* 633–638, 1979.

7. Abraham, G. E. The use of diatomite microcolumns for the chromatographic separation of steroids prior to radioimmunoassay. Pathol. Biol., *23:* 889–893, 1975.

8. Cassidenti, D. L., Vijod, A. G., Vijod, M. A., Stanczyk, F. Z., and Lobo, R. A. Short-term effects of smoking on the pharmacokinetic profiles of micronized estradiol in postmenopausal women. Am. J. Obstet. Gynecol., *163:* 1953–1960, 1990.

9. Wright, K., Collins, D. C., Musey, P. I., and Preedy, J. R. A specific radioimmunoassay for estrone sulfate in plasma and urine without hydrolysis. J. Clin. Endocrinol. & Metab., *47:* 1092–1098, 1978.

10. Loriaux, D. L., Ruder, H. J., and Lipsett, M. B. The measurement of estrone sulfate in plasma. Steroids, *18:* 463–472, 1971.

11. Abraham, G. E., Swerdloff, R. S., Tulchinsky, D., and Odell, W. D. Radioimmunoassay of plasma progesterone. J. Clin. Endocrinol. & Metab., *32:* 619–624, 1971.

12. Niswender, G. D., Menon, K. M., and Jaffe, R. B. Regulation of the corpus luteum during the menstrual cycle and early pregnancy. Fertil. Steril., *23:* 432–442, 1972.

13. Abraham, G. E., Maroulis, G. B., and Marshall, J. R. Evaluation of ovulation and corpus luteum function using measurements of plasma progesterone. Obstet. Gynecol., *44:* 522–525, 1974.

14. Niswender, G. D. Influence on the site of conjugation on the specificity of antibodies to progesterone. Steroids, *22:* 413–420, 1973.

15. Wu, C. H., and Lundy, L. E. Radioimmunoassay of plasma estrogens. Steroids, *18:* 91–111, 1971.

16. Stanczyk, F. Z., Shoupe, D., Nunez, V., Macias-Gonzales, P., Vijod, M. A., and Lobo, R. A. A randomized comparison of nonoral estradiol delivery in postmenopausal women. Am. J. Obstet. Gynecol., *159:* 1540–1546, 1988.

17. SAS Institute, Inc. SAS/STAT User's Guide, Version 6, Ed. 4, Vol. 2. Cary, NC: SAS Institute Inc., 1989.

18. Rao, C. R. Linear Statistical Inference and Its Applications, Ed. 2. New York: John Wiley & Sons, Inc., 1973.

19. Rosner, B., Willett, W. C., and Spiegelman, D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat. Med., *8:* 1051–1069, 1989.

20. Cauley, J. A., Gutai, J. P., Kuller, L. H., LeDonne, D., and Powell, J. G. The epidemiology of serum sex hormones in postmenopausal women. Am. J. Epidemiol., *129:* 1120–1131, 1989.

21. Toniolo, P., Koenig, K. L., Pasternack, B. S., Banerjee, S., Rosenberg, C., Shore, R. E., Strax, P., and Levitz, M. Reliability of measurements of total, protein-bound, and unbound estradiol in serum. Cancer Epidemiol., Biomarkers & Prev., *3:* 47–50, 1994.

22. Cauley, J. A., Gutai, J. P., Kuller, L. H., and Powell, J. G. Reliability and interrelations among serum sex hormones in postmenopausal women. Am. J. Epidemiol., *133:* 50–57, 1991.